Mineração de Dados em Biologia Molecular

Data Warehouse e OLAP

André C. P. L. F. de Carvalho
Monitor: Valéria Carvalho

Tópicos

- Data warehousing
- Data warehouse
- Sistemas de BD online
- OLAP
- Cubo de dados
- Operações com OLAP

Data Warehousing

- Processo de construir e usar data warehouses (armazéns de dados)
- Fornece estruturas (arquiteturas) e ferramentas necessárias para usuário organizar e analisar seus dados
  - Para tomada de decisões estratégicas

Data Warehousing

- Etapas necessárias:
  - Limpeza de dados
  - Integração de dados provenientes de diferentes BDs
  - Transformação dos dados
  - Consolidação de dados
    - Centralizar armazenamento dos dados integrados

Data Warehouse

- Importante passo de pré-processamento para MD
- Existem várias definições
  - Mas falta uma definição formal
- Uma BD de suporte à decisão que é mantida separadamente da BD da organização
  - Auxilia processamento de informação consolidando dados históricos para análise

Data Warehouse

Coleção de dados orientada a um tema, integrada, variante no tempo e não volátil para suporte ao processo de tomada de decisão

Orientada a um Tema

- Dados são organizado em torno de temas:
  - Ex.: pacientes, clientes, produtos, vendas
- Fornece uma visão simples e concisa sobre um tema
  - Excluindo dados que não são úteis no processo de tomada de decisão
- Foca em modelagem e análise de dados para tomadores de decisão
  - Não em operações diárias e processamento de transações

Integrada

- Integra dados de diversas fontes
  - Diferentes filiais, setores, empresas, centros de pesquisa, hospitais...
- Fontes são heterogêneas
  - BDs relacionais
  - Arquivos texto
  - Registros de transações online, etc

Integrada

- Para lidar com fontes diferentes, são aplicadas técnicas de limpeza e de integração
  - Garantem consistência dos dados
    - Convenções de nomes, formas de codificação, medidas de atributo, etc.
      - Formato para endereços em diferentes países
      - Profissão em diferentes regiões
      - Medidas de peso em diferentes países
      - Formato de dados de proteínas de diferentes repositórios
    - Dados são convertidos quando movidos para a data warehouse

Variante no Tempo

- Horizonte de tempo maior que BDs convencionais
  - Dados em um BD convencional armazenam apenas valores atuais
    - Ex.: dados de funcionários
  - Dados em uma DW varrem um período de tempo (por exemplo, 5 a 10 anos)
    - Toda estrutura utilizada como chave contém um elemento temporal
      - Explicita ou implicitamente

Não Volátil

- Após transformados, dados são armazenados em um novo local físico
- Atualização dos dados nos BDs não altera automaticamente dados armazenados na DW
  - Não precisa de mecanismos para controle de acesso simultâneo, recuperação e transação
    - Geralmente requer apenas duas operações: carregar e acessar os dados

Data Warehouse

- Armazém de dados semanticamente consistente
  - Serve como implementação física de um modelo de dados de suporte à decisão
- Arquitetura construída pela integração de dados de múltiplas fontes heterogêneas
  - Para suportar consultas estruturadas ou eventuais, relatórios analíticos e tomada de decisão

Data Warehouse

- A maioria das pessoas está familiarizada com SGBDs relacionais comerciais
  - Comparação com eles facilita entender funcionamento dos sistemas de DW

BD online x Datawarehouse

- Sistemas de BD online
  - Fornecem ferramentas que processam transações e consultas *online*
    - Cobrem operações do dia-a-dia de uma organização
      - Compras, controle de estoque, pagamentos, contabilidade, etc
  - O*n-line transaction processing systems* (OLTP)
- Datawarehouse
  - Fornecem ferramentas que auxiliam na análise de dados e tomada de decisão
    - Podem organizar e apresentar dados em vários formatos
      - Para satisfazer as necessidades diversas dos diferentes usuários
  - *On-Line Analytical Processing* (OLAP)

OLTP x OLAP

	OLTP	OLAP
Usuários	Profissional de TI, usuário	Analista de conhecimento
Função	Operações diárias	Suporte a decisão
Projeto de BD	Orientada a aplicação	Orientado a tema
Dados	Atuais, atualizados, tabela isolada	Históricos, resumidos, integrados, multidimensionais, consolidados
Uso	Repetitivo	Eventual
Acesso	Ler/escrever Indexação/hash chave prim.	Várias explorações
Unidade de trabalho	Transações curtas e simples	Perguntas complexas
#Registros acessados	Dezenas	Milhões
#Usuários	Milhares	Centenas
Tamanho da BD	100MB-GB	100GB-TB
Métrica	Volume de transações	Volume de consultas

OLAP

- Desenvolvido pelo mesmo criador dos BDs relacionais
  - E. F. Codd
- BDs relacionais: armazenam dados em tabelas
- OLAPs: Usam uma representação matricial multidimensional

OLAP

- Representação matricial multidimensional
  - Já era usada em estatística e outras áreas
    - Ex.: em planilhas eletrônicas
  - Facilita aplicação de várias operações de exploração e análise de dados
    - Forte foco em análise interativa de dados
    - Fornece capacidades para visualização e geração de resumos estatísticos

Matrizes Multidimensionais

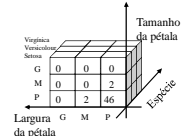
- Maioria dos conjuntos de dados são representados por tabela atributo-valor
  - Seja o conjunto de dados Iris

Tsépala	Lsépala	Tpétala	Lpétala	Classe
5.1	3.5	1.4	0.2	Setosa
4.9	3.0	1.4	0.2	Setosa
7.0	3.2	4.7	1.4	Versicol
7.6	3.0	6.6	2.1	Virginic
...	...	...	...	...

Conjunto Iris

- Iris (lírio): planta com flor
  - Atributos de entrada numéricos
    - Sepal length (cm)
    - Sepal width (cm)
    - Petal length (cm)
    - Petal width (cm)
  - Classes
    - Iris Setosa
    - Iris Versicolour
    - Iris Virginica
  - 150 exemplos, com distribuição 33/33/33

Matrizes Multidimensionais

- É fácil ver os mesmos dados utilizando uma matriz multidimensional
  - Tabelas podem ser convertidas em matrizes multidimencionais
    - Criação de matrizes multidimensionais

Criação de Matrizes Multidimensionais

- Passo 1: Identificar que atributos serão as dimensões e qual atributo será o atributo alvo
  - Atributos-dimensão (e valores), que definem as entradas da matriz multidimensional
  - Atributo alvo, que define o conteúdo das entradas ou células da matriz multidimensional
- Passo 2: Encontrar o valor para cada célula definida pelos valores das dimensões

Identificação dos atributos

- Atributos-dimensão devem assumir valores discretos e finitos
  - Uma dimensão para cada atributo
  - Número de valores de um atributo = variação de sua dimensão
  - Cada objeto é mapeado para uma célula na matriz multidimensional
    - Que pode representar vários objetos

Identificação dos Atributos

- O valor alvo é geralmente uma contagem ou valor contínuo
  - Ex.: custo de um item

Identificação dos Valores

- Como encontrar o valor de cada entrada da matriz?
  - Pela soma dos valores do atributo alvo nos exemplos com uma dada entrada ou
  - Pela contagem de todos os objetos que têm os mesmos valores dos atributos de entrada
    - Atributo alvo não necessariamente existe no conjunto original
- Gera uma tabela fato

Identificação dos Valores

- Conteúdo da entrada é a quantidade alvo que temos interesse em analisar
  - Ex: número de flores setosa (contagem) com tamanho da pétala e da sépala dentro de um certo limite

Exemplo

- Converter BD relacional do conjunto de dados íris para uma matriz multidimensional
  - Escolha dos atributos
    - Ex.: tamanho da pétala, largura da pétala e classe (espécie)
  - Conversão dos valores dos atributos numéricos para valores categóricos

Exemplo

- Conversão de valores numéricos para categóricos
  - Discretizar valores dos atributos de entrada para os valores categóricos **pequeno**, **médio** e **grande**
    - Exemplo:
      - *Lpétala: [0, 0.75), [0.75, 1.75), [1.75, ∞)*
      - *Tpétala: [0, 2.5), [2.5, 5), [5, ∞)*

Conjunto de Dados

Tsépala	Lsépala	Tpétala	Lpétala	Classe
5.1	3.5	1.4	0.2	Setosa
4.9	3.0	1.4	0.2	Setosa
7.0	3.2	4.7	1.4	Versicol
7.6	3.0	6.6	2.1	Virginic
...	...	...	...	...

Escolher Atributos

Tpétala	Lpétala	Classe
1.4	0.2	Setosa
1.4	0.2	Setosa
4.7	1.4	Versicol
6.6	2.1	Virginic
...	...	...

Discretizar

Tpétala	Lpétala	Classe
1.4	0.2	Setosa
1.4	0.2	Setosa
4.7	1.4	Versicol
6.6	2.1	Virginic
...	...	...

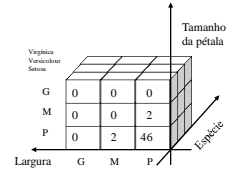
Tpétala
- pequeno:
- médio:
- grande:

Lpétala
- pequeno:
- médio:
- grande:

Tabela Fato

Tpétala	Lpétala	Classe	Contagem
pequeno	pequeno	setosa	46
pequeno	médio	setosa	2
médio	pequeno	setosa	2
médio	médio	versicolour	43
médio	grande	versicolour	3
médio	grande	virginica	3
grande	médio	versicolour	2
grande	médio	virginica	3
grande	grande	versicolour	2
grande	grande	virginica	44

Exemplo

- Cada tupla dos 3 atributos corresponde a um elemento da matriz
  - Esse elemento recebe o valor de contagem correspondente
  - As tuplas não especificadas recebem o valor 0

Exemplo

- Para facilitar a análise, usar fatias da matriz
  - Fatias podem ser exibidas como um conjunto de 3 tabelas bidimensionais
    - Uma para cada valor de um dos atributos classe ou espécie

Exemplo

- Fatias da matriz dimensional
  - O que é possível concluir a partir delas?

Setosa (linhas: Tamanho; colunas: Largura)

	pequeno	médio	grande
pequeno	46	2	0
médio	2	0	0
grande	0	0	0

Versicolour (linhas: Tamanho; colunas: Largura)

	pequeno	médio	grande
pequeno	0	0	0
médio	0	43	3
grande	0	2	2

Virginica (linhas: Tamanho; colunas: Largura)

	pequeno	médio	grande
pequeno	0	0	0
médio	0	0	3
grande	0	3	44

Exemplo

- Fatias da matriz dimensional
  - O que é possível concluir a partir delas?

Cada classe é caracterizada por combinações diferentes de valores

Setosa (linhas: Tamanho; colunas: Largura)

	pequeno	médio	grande
pequeno	46	2	0
médio	2	0	0
grande	0	0	0

Versicolour (linhas: Tamanho; colunas: Largura)

	pequeno	médio	grande
pequeno	0	0	0
médio	0	43	3
grande	0	2	2

Virginica (linhas: Tamanho; colunas: Largura)

	pequeno	médio	grande
pequeno	0	0	0
médio	0	0	3
grande	0	3	44

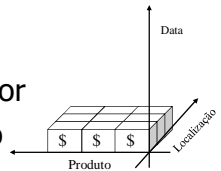
Exemplo 2

- Seja um conjunto de dados referentes a vendas de produtos
  - Identidade do produto
  - Localização do produto
  - Data
  - Valor

Conjunto de Dados

ProdutoId	Local	Data	Valor
1	São Carlos	18/10/2004	250
1	Bauru	1/9/2004	300
....	...	...	...
1	Araraquara	6/3/2004	430
...	...	...	...
27	Bauru	22/12/2004	200
27	Araraquara	7/10/2004	300
...	...	...	...
27	São Carlos	10/5/2004	500

Exemplo 2

- Supor que os seguintes atributos sejam selecionados para as dimensões
  - Identidade do produto
  - Localização
  - Data
- Atributo alvo é o valor
- Construir tabela fato

OLAP

- Em resumo:
  - Ponto inicial: conjunto de dados em formato de tabela
  - Para representar os dados como uma matriz multidimensional
    - Identificar as dimensões
    - Identificar o atributo alvo (foco da análise)
    - Construir tabela fato
    - Construir matriz multidimensional
    - Fatiar matriz

Exercício

- Dada a tabela ao lado:
  - Selecionar 1 subconjunto de 3 atributos-dimensão e 1 atributo alvo
  - Construir tabela fato
  - Mostrar matriz multidimensional
  - Mostrar fatias da matriz

Peso	*Altura*	*Idade*	*Salário*	*Atividade*
70	*168*	*40*	*2000*	*Professor*
58	*185*	*32*	*3500*	*Jogador*
85	*190*	*25*	*3000*	*Jogador*
60	*170*	*34*	*1000*	*Professor*
80	*165*	*37*	*1000*	*Professor*
65	*170*	*26*	*4500*	*Jogador*
90	*190*	*22*	*6000*	*Jogador*
49	*174*	*44*	*1300*	*Professor*
68	*188*	*30*	*3200*	*Professor*
75	*192*	*24*	*4000*	*Jogador*

OLAP

- Motivação chave para visão multidimensional dos dados:
  - Permitir agregar os dados de diferentes maneiras
  - Ex.: Dados de vendas
    - Vendas para um ano específico em uma dada cidade
    - Vendas anuais de um certo produto em uma dada cidade

OLAP

- Computação de totais agregados
  - Fixar valores para alguns dos atributos-dimensão
  - Usando esses valores fixos, agregar (somar) todos os valores para os demais atributos
    - Existem outros tipos de agregados além da soma

OLAP

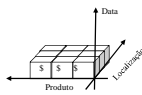
- Operação chave: formação de um cubo de dados
  - Estrutura multidimensional de dados junto com todas as possíveis agregações
    - Agregações que resultem da:
      - Seleção de um subconjunto adequado dos atributos-dimensão com junção das dimensões restantes

Exemplo

- Escolher a dimensão classe do conjunto íris e juntar as outras dimensões
  - Resultado:
    - Vetor com três elementos, cada um com o número de flores de cada tipo (classe)

Virgínica	Versicolour	Setosa
50	50	50

Exemplo 2

- Seja um conjunto de dados com vendas de vários produtos de várias empresas em diferentes datas
  - Pode ser representado por uma matriz 3-dimensional
  - Nessa matriz, existem:
    - 3 agregados 2-dimensionais
    - 3 agregados 1-dimensionais
    - 1 agregado 0-dimensional (total geral)

Data
$ $ $
Produto
Localização

Exemplo 2

- Um dos agregados 2-dimensionais
  - Junto com 2 agregados de 1 dimensão e o total geral (0-dimensional)

product ID \ date	Jan 1, 2004	Jan 2, 2004	…	Dec 31, 2004	total
1	$1,001	$987	…	$891	$370,000
⋮	⋮			⋮	⋮
27	$10,265	$10,225	…	$9,325	$3,800,020
⋮	⋮			⋮	⋮
total	$527,362	$532,953	…	$631,221	$227,352,127

Exemplo 2

- Um dos agregados 2-dimensionais
  - Junto com 2 agregados de 1 dimensão e o total geral

product ID \ date	Jan 1, 2004	Jan 2, 2004	…	Dec 31, 2004	total
1	$1,001	$987	…	$891	$370,000
⋮	⋮			⋮	⋮
27	$10,265	$10,225	…	$9,325	$3,800,020
⋮	⋮			⋮	⋮
total	$527,362	$532,953	…	$631,221	$227,352,127

Exemplo 2

- Um dos agregados 2-dimensionais
  - Junto com 2 agregados de 1 dimensão e o total geral

product ID \ date	Jan 1, 2004	Jan 2, 2004	…	Dec 31, 2004	total
1	$1,001	$987	…	$891	$370,000
⋮	⋮			⋮	⋮
27	$10,265	$10,225	…	$9,325	$3,800,020
⋮	⋮			⋮	⋮
total	$527,362	$532,953	…	$631,221	$227,352,127

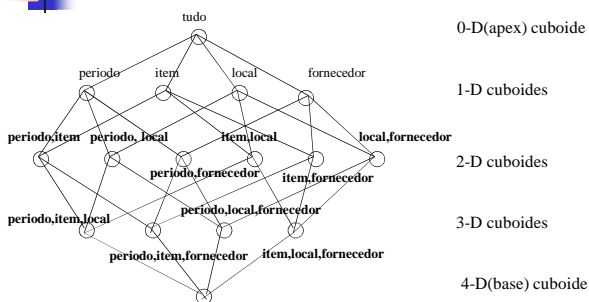
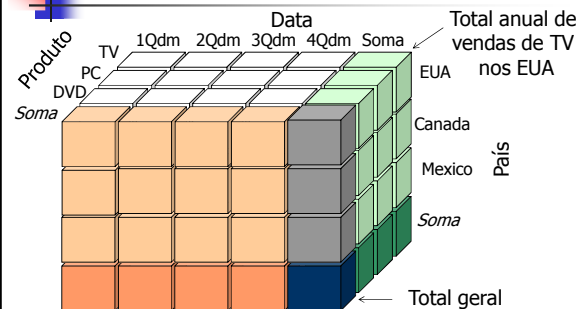
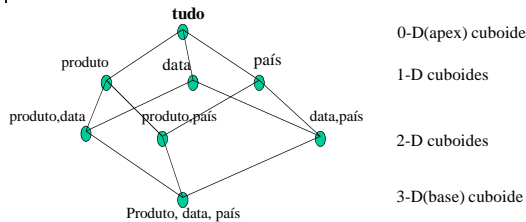
product ID \ date	Jan 1, 2004	Jan 2, 2004	...	Dec 31, 2004	total
1	$1,001	$987	...	$891	$370,000
⋮	⋮			⋮	⋮
27	$10,265	$10,225	...	$9,325	$3,800,020
⋮	⋮			⋮	⋮
total	$527,362	$532,953	...	$631,221	$227,352,127

Peso	Altura	Idade	Salário	Atividade
70	168	40	2000	Professor
58	185	32	3500	Jogador
85	190	25	3000	Jogador
60	170	34	1000	Professor
80	165	37	1000	Professor
65	170	26	4500	Jogador
90	190	22	6000	Jogador
49	174	44	1300	Professor
68	188	30	3200	Professor
75	192	24	4000	Jogador

Cubo de dados

- Permite que dados sejam modelados e vistos e várias dimensões
- Apesar do, tamanho das dimensões não precisa ser o mesmo
  - Pode ter menos ou mais que 3 dimensões
- Generalização do que é conhecido em estatística como tabulação cruzada

Cubo de dados

- Conhecido como cuboide de dados
- Hierarquia (reticulado) de níveis
  - Topo: O-D cuboide (cuboide apex)
    - Sumarização dos dados
  - Base: n-D cuboide (cuboide base)
- Reticulado de cuboides forma cubo de dados

Operações OLAP

- Agregação pode ser vista como uma forma de redução de dimensionalidade
  - Elimina uma ou mais colunas (linhas) somando seus valores
  - Uma coluna (linha) de células vira uma única célula
  - Exemplo mostrado para os dados de vendas reduziu de 3 para 2 dimensões

Pivotagem

- Agregar sobre todas as dimensões exceto duas

product ID	date: Jan 1, 2004	Jan 2, 2004	...	Dec 31, 2004	total
1	$1,001	$987	...	$891	$370,000
⋮	⋮			⋮	⋮
27	$10,265	$10,225	...	$9,325	$3,800,020
⋮	⋮			⋮	⋮
total	$527,362	$532,953	...	$631,221	$227,352,127

Slicing

- Fatiamento
- Seleciona um grupo de células da matriz
  - Especificando um valor específico para uma ou mais dimensões

Largura

Tamanho	pequeno	médio	grande
	46	2	0
	2	0	0
	0	0	0

Setosa

Dicing

- Cortar em cubos
- Seleciona um subconjunto de células da matriz
  - Especificando um intervalo de valores para atributos
  - Equivalente a definir uma sub-matriz da matriz completa

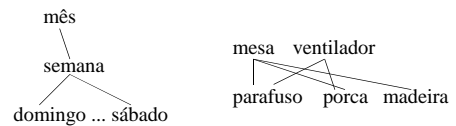
ProdutoId	Local	Valor
1	São Carlos	250
1	Bauru	300
....	...	...
1	Araraquara	430
...	...	...

Slicing e Dicing podem ser acompanhados por operações de agregação sobre algumas dimensões

Roll-Up e Drill-Down

- Cada atributo não precisa ser visto como atômico
  - Pode ter propriedades associadas a ele
    - Data
      - Ano
      - Mês
      - Dia
      - Semestre
    - Local
      - Continente
      - País
      - Estado
      - Cidade
    - Produto
      - Roupa
      - Móveis
      - Peças
      - Alimentos

Roll-Up e Drill-Down

- Categorias podem ser organizadas em redes ou árvores hierárquicas

Roll-Up e Drill-Down

- Roll-up
  - Subir na hierarquia
  - Agregar valores do atributo
  - Ex.: Vendas do mês somando vendas do dia
- Drill-down
  - Descer na hierarquia
  - Quebrar valores do atributo
  - Ex.: Vendas diárias quebrando as vendas do mês

Roll-Up e Drill-Down

- Diferentes granularidades devem estar disponíveis na tabela
- Relacionadas a agregação
  - Mas fazem várias agregações de células dentro de uma dimensão
  - e não uma única agregação através da dimensão inteira

Observações

- Existem outros tipos de sistemas de BDs que suportam análise de dados multidimensionais
  - Alguns são baseados em BDs relacionais
    - Conhecidos como sistemas ROLAP
  - Alguns usam como modelo de dados um representação especificamente multidimensional
    - Conhecidos como sistemas MOLAP
  - BDs estatísticos também foram desenvolvidos para armazenar e analisar vários tipos de dados estatísticos
    - Conhecidos como SDBs

Exercício

- Dada a tabela ao lado:
  - Mostrar pelo menos um exemplo para as operações de:
    - Dicing
    - Slicing
    - Roll-Up
    - Drill-Down

Peso	*Altura*	*Idade*	*Salário*	*Atividade*
70	*168*	*40*	*2000*	*Professor*
58	*185*	*32*	*3500*	*Jogador*
85	*190*	*25*	*3000*	*Jogador*
60	*170*	*34*	*1000*	*Professor*
80	*165*	*37*	*1000*	*Professor*
65	*170*	*26*	*4500*	*Jogador*
90	*190*	*22*	*6000*	*Jogador*
49	*174*	*44*	*1300*	*Professor*
68	*188*	*30*	*3200*	*Professor*
75	*192*	*24*	*4000*	*Jogador*

Considerações Finais

- Data warehousing
- Data warehouse
  - Armazena grandes quantidades de dados
  - Pode ser dividida em unidades lógicas menores (data marts)
- OLAP
- Cubos de dados
- Operações com OLAP